

**UNITED STATES DISTRICT COURT  
FOR THE DISTRICT OF MASSACHUSETTS**

THE HIPSAVER COMPANY, INC., Plaintiff / Counterclaim Defendant,	)	
	)	
	)	
v	)	Civil Action No. 05-10917 PBS
	)	
J.T. POSEY COMPANY,	)	
Defendant / Counterclaim Plaintiff.	)	
	)	
	)	

**HIPSAVER’S MEMORANDUM IN SUPPORT OF ITS MOTION *IN LIMINE*  
TO EXCLUDE THE 2006 TAMPERE TEST AND TEST RESULTS**

Plaintiff, the HipSaver Company, Inc. (“HipSaver”) respectfully submits this memorandum in support of its motion *in limine* to exclude a test and test results conducted at Tampere Laboratories in the Spring of 2006 (“the Tampere Test”), attached as **Exhibit A**. Defendant J.T. Posey (“Posey”) seeks to admit test results from the Tampere Test in order to support claims regarding its Hipster product made in advertisements from as early as 2001. The Tampere Test is inadmissible for such purpose because (a) the Tampere Test results are irrelevant to the falsity of Posey’s establishment claims because they were not cited in the accused advertisements and moreover, did not exist at the time of the advertisements; and (b) the Tampere Test, even if relevant, is inadmissible hearsay under the Federal Rules of Evidence. Accordingly, for the reasons explained below, the Tampere Test should not be admitted.

## FACTS

This lawsuit concerns the literally false advertising of Defendant J.T. Posey. Posey ran an advertising campaign from 2001 through 2005 that touted Posey Hipster's performance in a fall. Specifically, one of the advertisements proclaims that "Posey Hipsters Help Protect Against Injury From Falls" and provides a graph below the statement claiming to show the force attenuation abilities of a Posey Hipster in a test simulating a fall (herein referred to as the "Garwood Advertisement"). *See* Garwood advertisement in black and white attached at **Exhibit B** (HS2 000063). In different variations, the Garwood Advertisement refers to these impact tests as a "...test performed by Garwood Laboratories" or as "Independent Laboratories testing..." Posey never claimed that the claims and underlying tests set forth in the advertisements were based on any research by Tampere University or any other research facility.

The challenged Garwood Advertisements containing these statements are comprised of a box of text that describes a test simulating a fall that was conducted by Garwood Laboratories in 2001. This text takes up about 30% of the advertisement and is often placed on the lower third of the page. On the top of some of these ads, in large font, a statement reads, "Posey Hipsters Help Protect Against Injury From Falls." This statement at the top of the page is clearly linked to the statements in the lower third of the page and all statements appear to rely on testing conducted at Garwood Laboratories in 2001. *See* **Exh. B**.

Because HipSaver believes that the claims in the Garwood Advertisement are false, HipSaver instituted the present action. Throughout the false advertising lawsuit, Defendant Posey asserted that the claims made in the Garwood advertising and the

underlying results presented in the advertising were based on testing conducted by Garwood Laboratories in July 2001. Through expert discovery, Posey's biomechanical expert, Dr. Ebramzadeh, agreed with HipSaver's expert that the Garwood test did not simulate a fall:

Dr. Ebramzadeh: [HipSaver's] objections were two categories. One category objected to the application of this type of test to simulate a fall. *I agree that this is not the proper way to simulate a fall.*

Deposition of Edward Ebramzadeh, November 10, 2006, p. 51, ll. 3-6.

Because the Garwood Tests did not simulate a fall and therefore, cannot support claims of the Posey Hipster garment's performance during a fall, Posey sought supplemental evidence. In the Spring of 2006, almost one year after the present law suit was filed, Posey contracted Dr. Jari Parkkari at Tampere University in Finland to conduct impact tests on foams used in hip protector products. See **Exh. A**.

The Tampere Test was commissioned solely in preparation for the present litigation and was intended to supplement the Garwood testing cited in the Garwood Advertisement. In the 2006 Tampere Test, an impact study was conducted that was intended to simulate a fall, using a surrogate pelvis and/or femur. The Tampere Test and results were not published and have not been subject to peer review. In fact, Posey's biomechanical expert, Dr. Ebramzadeh, was not even personally involved in the 2006 Tampere Test and has no personal knowledge of the Tampere Test, the Tampere Test protocol or the results of the Tampere Test.

### **ARGUMENT**

The 2006 Tampere Test is irrelevant and inadmissible for the purpose of supporting Posey's establishment claims, and particularly to those claims relating to the

simulation of a fall because (a) the 2006 Tampere Test results were not cited in the accused advertisements and moreover, did not exist at the time of the advertisements and therefore, could not have been relied upon to make the claims contained in the Garwood Advertisement; (b) the Tampere Test, even if relevant, is inadmissible hearsay because the commissioned study was not published or subject to peer review and will be offered through Posey's expert, who has no personal knowledge of the Tampere Test. The Tampere Test simply cannot be used to support the claims for which it is intended to be offered.

**A. Posey's Post-Litigation Tampere Test is Irrelevant to Any Assessment of the Falsity of the Statements Contained in the Garwood Advertisement Because the Tampere Tests Did not Exist at the Time of the Garwood Advertisements and Could Not Have Supported the Claims at the Time They were Made**

**1. Only the Garwood Tests are relevant to a determination of the falsity of Posey's establishment claims.**

Because an establishment claim is literally false if the cited tests do not establish the proposition for which they are cited, only the cited Garwood Test (which was cited in the Garwood Advertisement), and not the Tampere Test (which did not even exist at the time of the Garwood Advertisement), is relevant to this litigation. *Castrol, Inc. v. Quaker State Corp.*, 977 F.2d 57, 63 (2d Cir. 1992) cited by the Honorable Judge Lindsay in *Gillette Company v. Norelco Consumer Products Company*, 946 F. Supp. 115, 121-122 (D. Mass. 1996). The challenged Garwood Advertisements are comprised of a box of text that describes a test simulating a fall that was conducted by Garwood Laboratories in 2001. This text takes up about 30% of the advertisement and is often placed on the lower third of the page. On the top of some of these ads, in large font, a statement reads, "Posey Hipsters Help Protect Against Injury From Falls." This statement at the top of the

page is clearly linked to the statements in the lower third of the page and all statements rely on testing conducted at Garwood Laboratories in 2001. Only the 2001 Garwood Test is cited in the Advertisement and therefore, only the Garwood Advertisement is relevant to a determination of falsity.

**2. The 2006 Tampere Test is Irrelevant to a Determination of the Falsity of Posey's Establishment Claim.**

The Tampere Test is irrelevant to the falsity of Posey's establishment claim because it was not cited, either explicitly or implicitly, in the accused advertisements and indeed, did not even exist at the time of the accused advertisements. In order to prove that an establishment claim is literally false under the Lanham Act, a plaintiff must prove that the tests cited to establish a product's performance or superiority in the advertisement at issue do not stand for the proposition for which they are cited: "[A] plaintiff can meet this burden by demonstrating that the tests were not sufficiently reliable to permit a conclusion that the product is superior." "If the plaintiff can show that the tests . . . do not establish the proposition asserted by the defendant, the plaintiff has obviously met its burden." *Castrol, Inc. v. Quaker State Corp.*, 977 F.2d 57, 63 (2d Cir. 1992) cited by the Honorable Judge Lindsay in *Gillette Company v. Norelco Consumer Products Company*, 946 F. Supp. 115, 121-122 (D. Mass. 1996); *see also Procter & Gamble Co. v. Chesebrough-Pond's, Inc.*, 747 F.2d 114, 119 (2d Cir. 1984).

Because the Tampere Test was not cited in the Garwood Advertisement, and indeed could not have been cited in the Garwood Advertisement since the test wasn't even conducted until almost a year after the advertisement was withdrawn, the Tampere Test is irrelevant to Posey's defense of its claim.

**B. The Tampere Test and Test Results are Inadmissible Under the Federal Rules of Evidence.**

The Tampere Test is a commissioned work presumably offered into evidence by Posey to prove that Posey padding can adequately reduce the impact force below the fracture threshold. The study was not published nor peer-reviewed. Even Posey's expert, Dr. Ebramzadeh, has no personal knowledge of the Tampere Test. Accordingly, the Test is inadmissible hearsay and should not be admitted. Fed. R. Evid. R. 801.

As stated above, the Tampere Test was commissioned by Posey and completed in 2006. Posey's expert, Dr. Ebramzadeh, cursorily referenced a test in Finland in his supplemental report (likely the Tampere Test) but never discussed the test or the results in any detail and did not produce the Test or any underlying material in his supplemental expert report. *See Exh. C*, Ebramzadeh Supplemental Expert Report at unnumbered p. 12. Furthermore, Dr. Ebramzadeh was not personally involved in the Tampere Test and has no personal knowledge of the Tampere Test. In fact, he could only state in his supplemental expert report that he "understand[s] that the Posey products tested in the [Tampere Test] used the same foam that was used at the time the Posey ad in question was published" and that he "understand[s] that one of the products tested used the same foam that was used in the Garwood test of July 2001." His "understanding" clearly came from a third source. Dr. Ebramzadeh does not have any personal knowledge of the Test and cannot authenticate the Test or Test results or speak to the validity or reliability of the test.

At best, a commissioned test such as the Tampere Test may only be admitted as admissible hearsay under Fed. R. Evid. 803 (18) if it published and if the publication is established as a reliable authority by the testimony or admission of the witness through

whom it is offered or by other expert testimony or judicial notice. Fed. R. Evid. 803 (18). But, the Tampere Test was not published and has not been peer reviewed. As discussed above, Dr. Ebramzadeh offers no personal knowledge of the Test or the Test protocol and cannot testify as to the reliability or authority of the Test. Accordingly, without any personal knowledge to authenticate the Test and without any ability to testify to the reliability or validity of the 2006 Tampere Test, the test cannot even be read into evidence.

### **CONCLUSION**

Accordingly, because the Tampere Test conducted in 2006 is irrelevant to this litigation and, even if relevant, is inadmissible under the Federal Rules of Evidence, the 2006 Tampere Test must be excluded.

THE HIPSAVER COMPANY, INC.  
By its Attorneys,

/s/ Courtney M. Quish  
Lee Carl Bromberg  
BBO No.: 058480  
Edward J. Dailey  
BBO No.: 112220  
Courtney M. Quish  
BBO No.: 662288  
BROMBERG SUNSTEIN, LLP  
125 Summer Street - 11th floor  
Boston, Massachusetts 02110-1618  
617.443.9292  
[cquish@bromsun.com](mailto:cquish@bromsun.com)

Dated: May 15, 2007

**CERTIFICATE OF SERVICE**

I certify that this document has been filed through the Electronic Case Filing System of the United States District Court for the District of Massachusetts and will be served electronically by the court to the Registered Participants identified in the Notice of Electronic filing.

/s/ Courtney M. Quish \_\_\_\_\_

May 15, 2007

02820/00502 664187.1




**Tampere University of Technology, Applied Mechanics**
**Jari Parkkari**
**Jarmo Poutala**
**P.O. Box 589**
**SF-33101 Tampere**
**Finland**
**E-mail:**
**jarmo.poutala@tut.fi**
**jari.parkkari@uta.fi**

## Hip protector test

A series of impact experiments were conducted to measure the force attenuation provided by Posey Company's hip pads typed 6016M and 6018HM for innerwear use. The tests were performed at the midrange force of 7550 N as per the protocol and the testing system described in *Bone 1999 Aug. 25(2):229-35*. The above-mentioned force was attenuated by soft tissue to the value of 5650 N, which match the average peak hip impact force measured in the muscle-relaxed state during in vitro falling tests (*Robinovitch et al. 1991*). Pad named PT-230 (thickness 20 mm) was used to simulate the soft tissue. This surrogate soft tissue was replaced after every impact.

The data acquisition system was based on Microstar Laboratories Data Acquisition Processor DAP 3200A (Bellevue, U.S.A). The DAP 3200A has the DAPL operating system. The sampling time was 10  $\mu$ s. The number of acquired points was 1500 for each test curve. The acquired data were analyzed by Matlab, which is used to numeric computation and visualization. The Matlab is a trademark of the Math Works (Natick, U.S.A). Known KPH Hip Protectors were used to see the same impact force level as reached in the tests earlier. The testing system is shown in Figure 1.

Pads were removed from the underpants to get them fixed and positioned in a standardized way for testing. The thicknesses of the tested hip pads were 12-12.5 mm. Six impact tests were conducted for each protector type (only one impact was allowed on one device). The force measurements were filtered and averaged peak values and standard deviations were calculated to get the maximum compressive impact forces.

The results of the experiment are shown in Table 1. Averaged time-dependent test curves of hip protectors are shown in Figure 2.

Typical test curves of Posey pads are shown in Figures 3 and 4. Peak values of impact forces of all Posey tests are seen in Tables 2-7. The pad number 6 of both models were tested six consecutive times to see possible degeneracy after impacts. The results of these repeated impacts are shown in Tables 3 and 6.

Two pairs of pants were kept in oven 3 hours at the small elevated temperature of 86 °F (30 °C) to see possible changes in their force attenuation capacity. The room temperature was 71.6 °F (22 °C). The elevated temperature test results are shown in Tables 4 and 7. Small decline in the force attenuation capacity of the pads was observed in elevated temperature compared to those impact tests conducted in room temperature (Table 1).

**PC 5737**

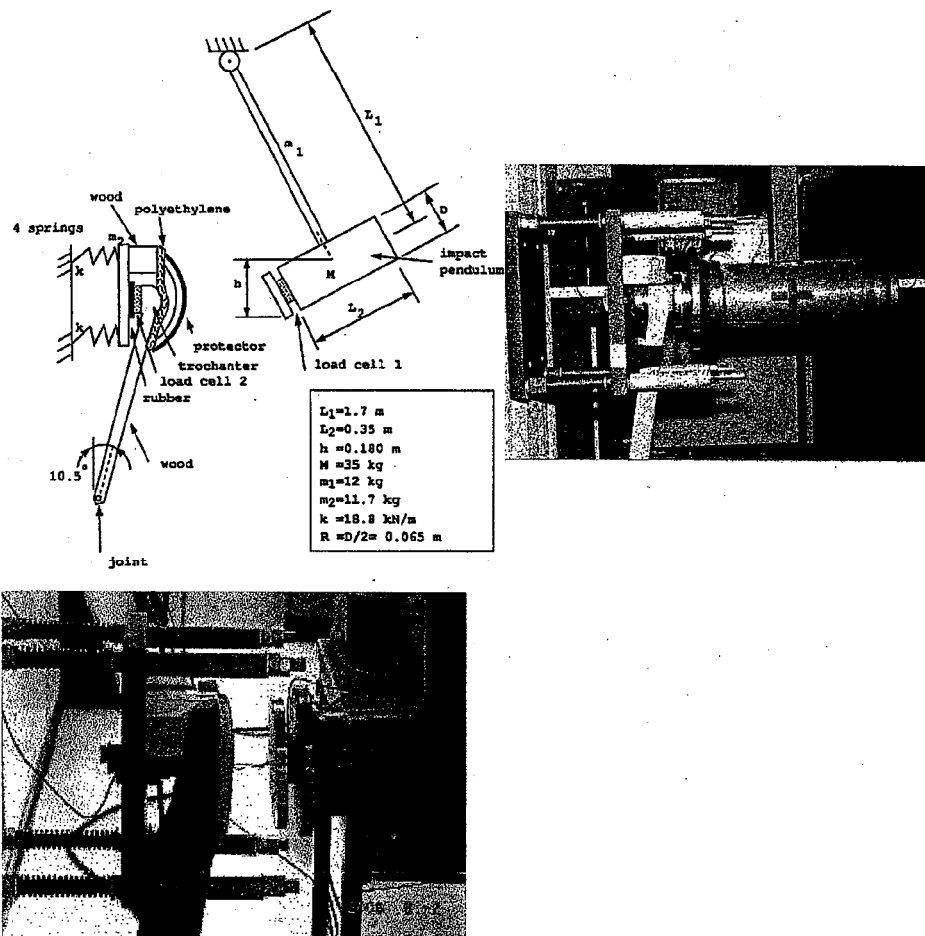
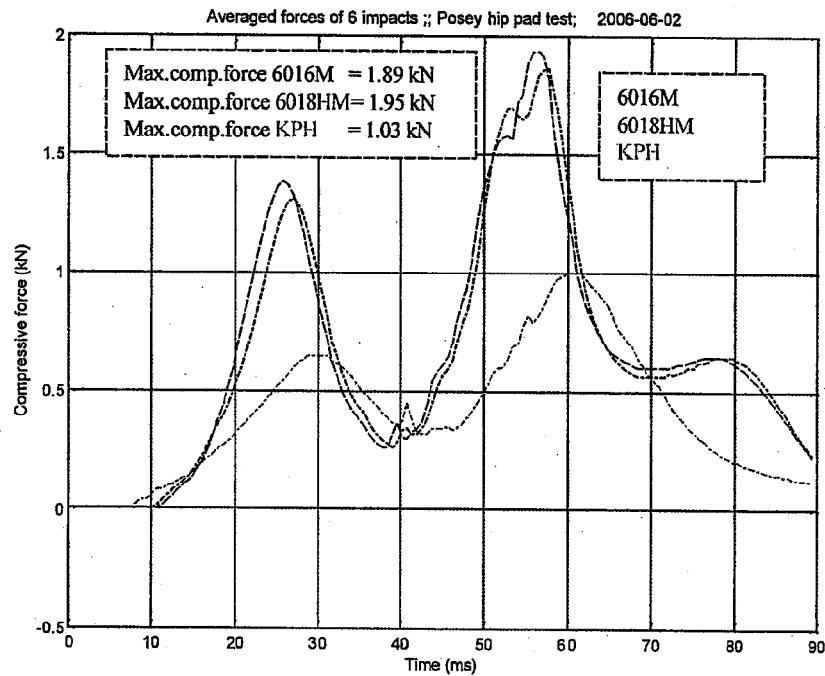


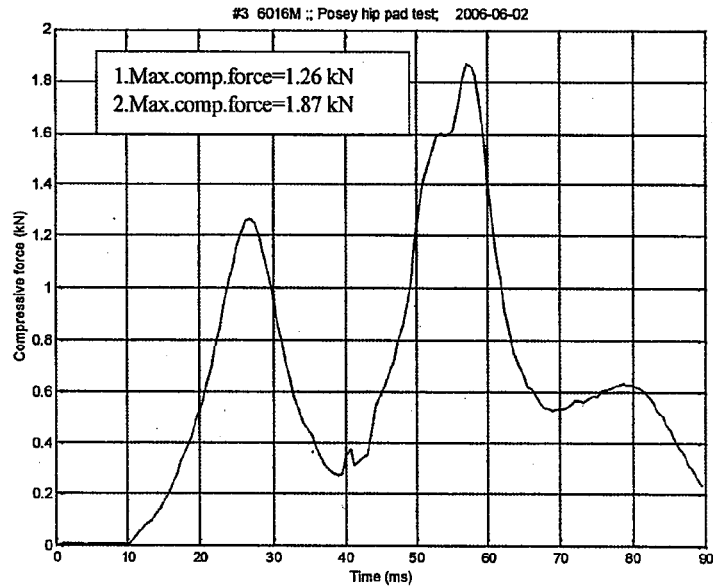
Figure 1 The hip protector testing system.



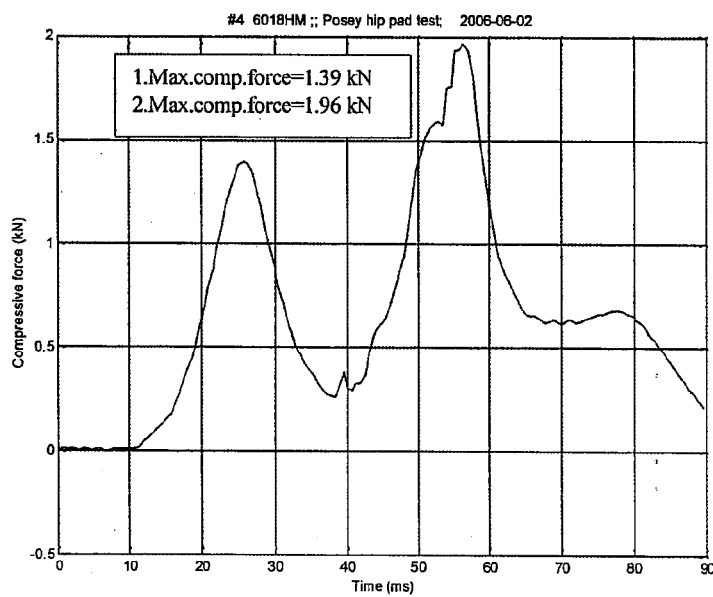
**Figure 2** Averaged test curves for hip protectors.

**Table 1** Averaged trochanteric impact forces and their standard deviations under the typical force created by a fall on the hip (5650 N).

Test	Speed m/s	Energy Nm	KPH		Posey 6016M #1,2,3,4,5,6		Posey 6018HM #1,2,3,4,5,6	
			Mean kN	Std kN	Mean kN	Std kN	Mean kN	Std kN
Six impacts	1.9	74	1.02	0.096	1.89	0.109	1.95	0.068



**Figure 3** Typical impact test curves of the Posey 6016M pad (#3).



**Figure 4** Typical impact test curves of the Posey 6018HM pad (#4)

PC 5740

**Table 2** Impact test values of six Posey 6016M pads (see Fig.3).

Pad #	Troc. peak 1 kN	Troc. peak 2 kN
1	1.46	2.07
2	1.33	1.95
3	1.26	1.87
4	1.26	1.76
5	1.22	1.83
6	1.31	1.85

**Table 3** Peak values of six repeated impacts of Posey 6016M pad # 6.

Impact #	Troc. peak 1 kN	Troc. peak 2 kN
1	1.31	1.85
2	1.50	2.30
3	1.53	2.11
4	1.57	2.19
5	1.53	2.29
6	1.64	2.27

**Table 4** Impact test values of four Posey 6016M pads at elevated temperature of 86 °F.

Pad #	Troc. peak 1 kN	Troc. peak 2 kN
8a	1.46	2.39
8b	1.48	2.25
9a	1.59	2.47
9b	1.55	2.44

**Table 5** Impact test values of six Posey 6018HM pads (see Fig.4).

Pad #	Troc. peak 1 kN	Troc. peak 2 kN
1	1.43	2.06
2	1.35	1.98
3	1.37	1.86
4	1.39	1.96
5	1.38	1.91
6	1.35	1.90

PC 5741

**Table 6** Peak values of six repeated impacts of Posey 6018HM pad # 6.

Impact #	Troc. peak 1 kN	Troc. peak 2 kN
1	1.35	1.90
2	1.37	1.96
3	1.36	2.11
4	1.34	1.97
5	1.37	2.04
6	1.35	1.92

**Table 7** Impact test values of four Posey 6018HM pads at elevated temperature of 86 °F.

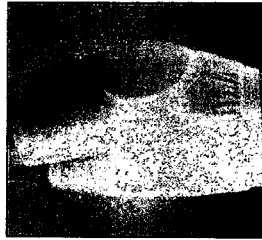
Pad #	Troc. peak 1 kN	Troc. peak 2 kN
8a	1.47	2.14
8b	1.49	2.01
9a	1.47	2.08
9b	1.48	2.18

Tampere 2006-06-21

Jarmo Poutala, Laboratory Manager

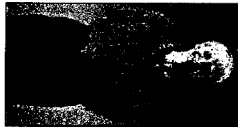
PC 5742

# POSEY HIPSTERS HELP PROTECT AGAINST INJURY FROM FALLS



It's a long way down for residents at risk of injury from falls. You can greatly reduce that risk with Posey Hipsters. The Hipsters' high energy-absorbing foam pads are positioned precisely over the hip bones, increasing the odds of surviving a fall uninjured. The Hipsters are comfortable and slim enough to be virtually undetectable under clothing. By offering increased protection, Hipsters relieve residents' anxiety about falling and enhance their quality of life.

- High impact-absorbing viscoelastic pads protect hip bones against injury from falls
- Soft, comfortable pads improve compliance versus hard-shelled products
- Washable to CDC standards for soiled linen without removing the pads
- 100% latex-free
- Five sizes for correct fit
- Discreet, low-profile pads are virtually undetectable under clothing

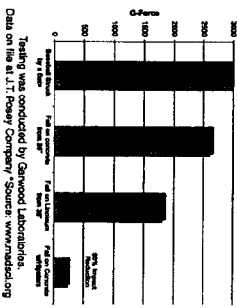


**Low Profile.**  
All styles fit discreetly under men's and women's clothing.

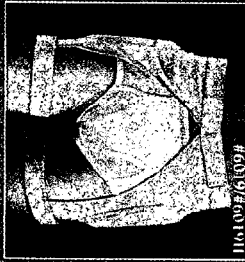
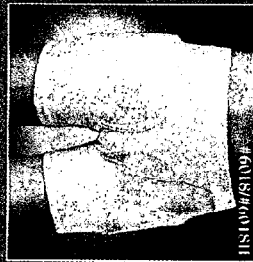
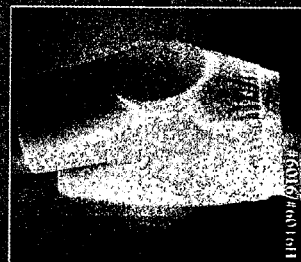


## Posey Hipsters Proven Effective in Laboratory Test

Posey engaged Garwood Laboratories to conduct testing to select a comfortable and effective impact absorbing material. A test was created that would simulate a fall causing direct impact to the greater trochanter. In this study, a weight was released in a guided drop to simulate a 120 lb subject falling from a height of 36", or the estimated height of the hip above the floor for a typical nursing home resident. The baseline measurement of impact force was determined to be a fall directly onto concrete. The G-force of a fall under this scenario was 2,600G's and, for purposes of comparison, is just slightly less impact force than that of a baseball being struck by a bat. In this extreme test, the low profile Posey Hipster reduced the impact force on average by 90% and showed excellent impact energy absorption.



**Special offer: 30-day no-risk, free trial.**  
Test the Posey Hipsters for yourself with no obligation to buy.



### POSEY #6016 HIPSTERS STANDARD BRIEF

- Easily fits over undergarments, or can be worn as underwear.
- Unisex sizing.
- #6016H Standard Brief with high durability pads.

### POSEY #6017 INCONTINENT BRIEF

- Snap front for easier application over diaper. Unisex sizing.
- #6017H Incontinent Brief with high durability pads.

### POSEY #6018 MALE FLY BRIEF

- Easily fits over undergarments, or can be worn as underwear.
- Fly front for improved compliance in male residents.
- #6018H Male Fly Brief with high durability pads.

### POSEY #6019 EZ-ON BRIEF

- Residents can wear their own undergarments.
- Can be worn in the shower.
- Hip pads can be removed for laundering or replacement.
- #6019H EZ-ON Brief with high durability pad.

Size	Waist Measurement	Hip Measurement
S	28" - 30" or 71 - 76cm	35" - 37" or 88 - 93cm
M	30" - 34" or 76 - 86cm	37" - 41" or 93 - 104cm
L	34" - 38" or 86 - 96cm	41" - 45" or 104 - 114cm
XL	38" - 42" or 96 - 106cm	45" - 49" or 114 - 124cm
XXL	42" - 46" or 106 - 116cm	49" - 53" or 124 - 134cm

Posey High Durability Hipsters contain denser foam than the standard Hipsters. This increased density aids in its ability to withstand higher heat washing and drying cycles.

### LAUNDERING INSTRUCTIONS:

- Hipsters**
  - WASH HOT
  - BLEACH AS DIRECTED ON CONTAINER
  - IRON ON
- High Durability Hipsters**
  - WASH HOT
  - BLEACH AS DIRECTED ON CONTAINER
  - IRON ON

J.T. Posey Company  
Arcadia, CA 91006 USA  
Tel: 800-447-6739  
www.posey.com

# EXHIBIT C

## PART 1



**UNITED STATES DISTRICT COURT**

THE HIPSAVER COMPANY, INC.,

Plaintiff,

V.

J.T. POSEY COMPANY,

Defendant.

AND RELATED COUNTERCLAIM.

Civil Action No. 05-10917 PBS

**REBUTTAL EXPERT REPORT OF EDWARD EBRAMZADEH  
PURSUANT TO  
RULE 26(a)(2)(B) OF THE FEDERAL RULES OF CIVIL PROCEDURE**

**CONFIDENTIAL – ATTORNEYS’ EYES ONLY**

## **I. INTRODUCTION**

In connection with my work in this matter, I was asked to review the Expert Summary of Opinions prepared by Dr. Wilson Hayes, and to express my opinions in response. My initial report on this matter was submitted on February 16, 2006. In preparing this report, I have been provided with some additional information and I have conducted interviews and phone conversations and have obtained additional literature and material. Additionally I have prepared some exhibits which are attached.

## **II. DATA AND OTHER INFORMATION RELIED UPON**

The data and other information I relied upon in forming my opinions in this report are listed in the attached Supplemental Appendix "A."

## **III. SUPPLEMENTAL OPINIONS**

Notwithstanding the differences between the Garwood test and the protocol test specified by ASTM F355-95, or the peer reviewed literature cited by Dr. Hayes in his report, in my view, the Garwood test (report dated August 2001) describes a test that can adequately be used to rank different materials by their ability to absorb impacts.

In the sections that follow, I will address various points raised by Dr. Hayes in his report.  
Page 5, Item 16:

The ASTM F 355-95 (hereafter called the ASTM standard or ASTM) states that this test should not be used, without some modifications, to test finished products. I understand that, in the Garwood tests, Posey evaluated a combination of raw materials which it was considering for use in its products and hard and soft finished products already on the market. The raw materials consisted of various foams and were prepared by cutting a specimen to a size generally slightly

larger than the 6 inch diameter impact surface and then encasing it in a protective pouch. The soft protective pads from finished products were prepared by cutting them from the underwear and placing them on the base of tower. Hard protective pads, such as the SafeHip, were laid on the base as they were. The dimensions of the pads were not always greater than the 6 inch diameter of the impact device; however, the contact area was approximately as large as or larger than the impact surface in most cases. Based upon my understanding, Garwood's specimen preparation, though not exactly according to ASTM F 355-95, would not have significantly affected the results of the comparisons of the various foams and soft pads being evaluated. However, in evaluating the results of the hard protectors, the shape of the protector has to be taken into account.

Page 6, Item 17:

The ASTM standard F 355-95 states that the mass of the impacted base should be at least 100 times that of the missile. This is specified in part to prevent ringing or vibration upon impact, which would make it difficult or impossible to take accurate measurements of impact.

I understand that Garwood conducted several sets of tests for Posey. In the first set of tests (report dated May 2, 2001), three different missiles, weighing 14 lbs, 28 lbs, and 55 lbs, to which the pads were fixed, were dropped from a height of 36 inches. This set of tests proved problematic because the missile would not always land in the desired position or orientation and, equally importantly, generated substantial amounts of vibration (ringing) in some cases, making accurate measurements impossible. Garwood addressed these issues by conducting a second set of tests using a guided drop tower, as described below.

In the second set of tests (report dated August 7, 2001), a guided missile weighing 72 lbs was used, and dropped from a height of 24 inches. Compared to the free fall used in the first set of tests, using the guided drop tower proved to substantially reduce ringing and vibration, and also allowed accurate and reproducible landing of the missile on the base. This guided drop tower was used in both the second set of tests (report dated August 7, 2001) and in subsequent tests.

I understand that only the data from the second set of tests, conducted with the guided missile (report dated August 7, 2001) was relied upon by Posey for conclusions about the materials, or for statements made in subsequent advertisements. In other words, due to the vibrations and ringing, the first set of tests was considered invalid and unreliable and was not used. Therefore, any objections with regard to ringing and vibrations were already addressed in the July 2001 tests, and the results should not be considered unreliable due to this issue.

It is important to note that, for the second (and subsequent) tests, Garwood used an industrial and commercially available impact tester (model ED50, manufactured by Turbo Reset), along with a missile that was built-in and designed for impact testing. Although the mass of the impacted base was not reported, the metal base of the machine was an integral part of the testing apparatus and it is reasonable to assume that the impact tester provided adequate stiffness and a base designed to sufficiently protect against vibration during impact. This is consistent with the fact that Garwood was successful in substantially reducing vibrations in the second tests.

Dr. Hayes estimated in his report that the sampling rate in the Garwood test was 16000 Hz, which is the minimum sampling rate specified in the ASTM F 355-95 to effectively measure impulses of up to 500 G. The greatest acceleration measured in the Garwood tests of July 2001 was 2662 G for the baseline measurement when no padding was used. Dr. Hayes pointed out that this sampling rate was inadequate for this particular impulse. However, it should be pointed out that most of the padding materials tested produced accelerations well below 500 G, which is within the range specified by ASTM, and none produced acceleration above 1200 G. Therefore, the sampling rate was adequate to measure impact response for most of the specimens.

The concern regarding sampling rate is based on the fact that, when the sampling rate is too low, the impact acceleration may be underestimated. However, I understand that at least insofar as the Garwood test is concerned, the materials that tested the highest were automatically eliminated from consideration by Posey. Therefore the fact that some of the materials tested exhibited acceleration above 500 G, is irrelevant. That is, for the remainder of the samples, which attenuated the acceleration to below 500 G, this limitation in measurement would not have affected specimen ranking and selection.

Page 6, Item 20:

The issue of specimen size and preparation was addressed in response to Item 17.

Page 6, Item 22:

The ASTM standard F 355-95 states that the time interval between consecutive tests of the same specimen should be  $3 \pm 0.25$  minutes. Dr. Hayes pointed out that, in the Garwood tests of July 26, 2001, the intervals between consecutive drops varied from less than one minute to

more than five minutes. I agree that this deviation from the protocol affected the results but, as discussed below, relative to the effect of material, this deviation is very small and does not affect the overall conclusions to be made regarding differences in impact absorption of the materials.

Supplemental Exhibit 1 shows the maximum acceleration for each of the three drops for each specimen. Each drop is shown in a different color bar. As this graph indicates, the differences among the three drops of the same specimen are generally much smaller compared to the differences from one material type to another. It should be kept in mind that other factors besides the time interval, such as fatigue, also could have affected variations from one drop to another. Therefore, it should not be expected that different drops of the same sample would be identical even if the time interval were held constant. Generally speaking, in this case, the differences in acceleration among the three runs are small. The greatest exception to this is SafeHip (number 19) run number 3, which is noticeably larger than the first two runs of the same. This may have been a result of structural failure of the hard shell during the first and second impacts, which would make it behave as a flat specimen in the third run. However, no special observations appeared in the documentation to indicate that there were any failures.

I conducted two statistical analyses similar to those conducted by Dr. Hayes to assess the effect of the time interval on the outcome (acceleration). A general linear model was established, with the acceleration as the dependent variable. The analysis was conducted two different ways. First, as in Dr. Hayes' analysis, material and drop (second or third) were introduced as categorical variables (factors). This analysis determined the relative effects of 1) drop (second or third repeat) and 2) material (specimen) on the impact acceleration. In the second analysis, the time interval was introduced as a covariate. This analysis determined the relative effects of 1) time interval and 2) material on impact acceleration.

The first analysis indicated that, although it is quite certain that there was a difference between drop 2 and drop 3, this difference was much smaller than the effect of material on impact acceleration. Likewise, the second analysis indicated that, although it is quite certain that there was an effect of time interval, this effect was much smaller than the effect of material on impact acceleration. This was indicated by the fact that the Type III sum of squares was three orders of magnitude larger for the variable specimen than for the variable time. This is one parameter that is related to the size of an effect. Additional parameter estimates were also consistent with this observation.

To understand these statistical analyses, it is important to distinguish between statistical significance and the size of difference. The P-value (which was used by Dr. Hayes in reaching his conclusion) is only one outcome of such an analysis. A P-value smaller than 0.05 is commonly considered to indicate a 'statistically significant' result. The smaller the value, the smaller the probability that the result was obtained by random chance. Therefore, the P-value is related to the certainty associated with a given observation, and not with the size. That is, the P-value alone does not say anything about how much of a difference is made.

It is well-established among the scientific community that it is incorrect to use the P-value alone in assessing differences observed in test. Rather, a thorough analysis requires an examination of the raw data, the actual differences observed in the samples, and the indicators of the size of the effect, such as the Type III sum of squares in this case, in addition to the P-value. (Supplemental Exhibit 2 discusses of the common misinterpretations of P-values in the literature.)

In summary, I do not agree that the fact that there was variation in the timing between drops made the results unreliable.

Page 7, Item 23:

The Garwood test can be viewed as a materials test intended to rank and select from among several different candidate materials for padding in a hip protector device (with special consideration required for the hard shells).

In his analysis, Dr. Hayes criticizes the Garwood tests conducted in March and April, 2001 for various reasons. However, as indicated, I understand that the first set of tests, that is, tests conducted in March and April of 2001, was considered by Posey unsuccessful and none of the results of these tests were used to draw any conclusions about the materials that were tested later, in July 2001. Nor were the results used in any advertising. Moreover, as discussed above, the vibrations and ringing that were present in the first set of tests, were addressed in the second set of tests. Also, the only test that used additional padding (linoleum, carpet, etc.) other than the material being tested was the first set of tests which, again, were abandoned by Posey because the data was considered unreliable. In all subsequent tests, there was no material between the padding and the base. Therefore, in my opinion, Dr. Hayes's objections indicated in this section do not make the statements made in the Posey ad unreliable.

Page 8, Item 26:

The tests conducted by Garwood do not model the intricacies of the bone and soft tissue geometries and material properties. However, in my opinion, the Garwood set of tests (report dated August 2001) is valid for comparison of shock attenuation of the materials tested. Moreover, despite the deviations from the ASTM specifications, the results of this set of tests are



sufficient to draw reasonably reliable conclusions regarding the relative shock attenuation properties of the materials tested.

Page 8-9, Items 26-28:

A load applied to a structure is distributed throughout the components of the structure as stresses. Put differently, stress is the internal distribution of forces within a component that react to the external loads applied to the component. Stress analysis is the science of determining the stresses in a component as a function of applied loads. Stresses are difficult to calculate unless the loads are simple and the geometry of the component is simple (e.g., cylindrical rod, beam, etc.). At any point in a structure, stress is described as load per unit area, e.g., lbs/in<sup>2</sup>.

In studying the structural properties of components, a simulation of true loads on the structure is often necessary. In such cases, a simulation of the relevant components and properties of the structure is also necessary. For example, to simulate loads on the hip during a fall, as emphasized by Dr. Hayes, the experiment should be designed to approximate the loads, material properties, geometries and contact conditions of typical real life cases as closely as possible.

By contrast, in a materials test, it is common and in fact often desirable to use simplified geometries because this facilitates comparison of material on an "all else being equal" basis, and can allow the calculation of stresses and subsequent generalization of the findings to other shapes and applications.

In his analysis of the Garwood test, Dr. Hayes placed great emphasis on the magnitude of the impact forces. By basing his analysis on forces, rather than on stresses, Dr. Hayes has treated the Garwood test as a fall simulation. However, in Section 26 and elsewhere, Dr. Hayes has

stated that the Garwood test was not a valid simulation. Since the Garwood test is not a valid simulation, then it should not be analyzed as one. Rather, it should be viewed as a materials test and analyzed as one.

Treating the Garwood test as a fall simulation rather than a materials test exaggerates the difference in outcome between the Garwood test and the tests conducted by Kannus et al. or Parkkari et al., and others who have simulated a fall. For example, Dr. Hayes estimated in his report the impact force of the Garwood test, without any padding, at about 191,800 lbs, which he states is orders of magnitude greater than forces on the hip during a fall. This comparison is misleading because it fails to take into consideration that, as compared to an impact on the human hip (as modeled by the above authors), the impact forces in the Garwood test were distributed more uniformly and over a much larger surface area, resulting in contact stresses that were not nearly as large as suggested by Dr. Hayes.

It is important to emphasize that the baseline force 191,800 lb force was especially high because it was a drop in which no padding was used. With any of the materials tested, the impact force was reduced by several times. In the Garwood experiment, a flat metal surface, approximately 6 inches in diameter, was dropped on each pad. This led to a highly uniform stress distribution over the area of the pad. The pads were generally slightly larger than the surface of the metal, though in a few cases they were slightly smaller. Therefore, for the impact force of 191,800 lbs, the peak contact compressive stress would be  $6787 \text{ lbs/in}^2$ . For a pad that created an acceleration of 500 G, the stress would be  $1275 \text{ lbs/in}^2$ .

By contrast, in the models constructed by Parkkari and others, the area of contact between the pendulum and the pad was much smaller. Further, the greater trochanter is not flat. Rather, it is an ellipsoidal surface. The experiments by Kannus et al and others use 'medium'

impact forces in excess of 1500 lbs, which are attenuated by the soft tissue but are still over 1300 lbs. The true stresses in these experiments are difficult to calculate but, to demonstrate the point, assuming that the contact area at point of impact is  $3 \text{ in}^2$ , an impact force of 1300 lbs would result in stresses in excess of  $430 \text{ lbs/in}^2$ . (I say "in excess" because of the curvature for which, again, stresses would not be uniformly distributed). Although these stresses are still lower than the stresses generated in the Garwood test, they are not lower by a factor of 99 to 152, the ratios estimated by Dr. Hayes for the impact forces.

In his analysis of the baseball/bat analogy, Dr. Hayes likewise used impact forces without taking the contact area and contact stresses into consideration. As a consequence, the ratios of the impact forces between a baseball and a bat to the impact forces in fall simulations by Kannus et al and others imply exaggerated differences in impact.

The Posey ad at issue in this case stated, "... 2660 G's ... for purposes of comparison, is just slightly less impact force than that of a baseball being struck by a bat." As support for this statement, the ad cited the website [www.madsci.org](http://www.madsci.org) which, in turn, cites a 1989 textbook: "Physics, Second Edition" by Hans C. Ohanian, published by W. W. Norton and Company. I note that on page 33 this textbook lists the acceleration of a baseball being hit by a bat at  $30000 \text{ m/s}^2$  (approximately 3000 G's). Therefore the statement in the ad is supported by a standard physics textbook. Moreover, the website indicates clearly that "'g'-forces are really a measure of acceleration ..."

Dr. Hayes estimated the peak acceleration of a baseball hitting a bat at between 8930 G to 11,500 G's. The discrepancies between the acceleration of 3000 G's quoted in the Posey ad and in Ohanian's textbook, compared to the range estimated in Dr. Hayes' report are perhaps due to the fact that he calculated the accelerations for major league players. However, it is reasonable to

assume that the pitching and off bat speeds of the average baseball player would be lower and the contact times would be higher in comparison to major league players, resulting in lower accelerations.

In summary, in my opinion, the Garwood test, despite its limitations and the deviations from the ASTM protocol, is still reliable enough to draw basic conclusions about the shock resistance of the materials that were tested. That is, it is reasonable to use the results for comparison of various padding materials for shock attenuation.

Page 10, Item 29:

The title of the Posey ad at issue was "POSEY HIPSTERS HELP PROTECT AGAINST INJURY FROM FALLS." This statement is supported by recent findings from a test conducted in Finland by Parkkari and Poutala. In that test, the investigators used a simulation to measure the force attenuation by Posey's hip pads. The impact force was 1697 lbs, which was attenuated to 1270 lbs by the soft tissue simulation. I understand that the Posey products tested used the same foam that was used at the time the Posey ad in question was published. I further understand that one of the products tested used the same foam that was used in the Garwood test of July 2001. This pad attenuated the impact force on average to 425 lbs, which is below the average hip fracture load. Therefore, the statement in the Posey ad is not literally false.

Page 10, Item 30:

Based on the material provided to me, it does not appear that the protocol devised by Minns et al. (HS2 000311) has been established as a part of any standard protocol. In his correspondence of September 2005, in which Professor Minns presented the data on the Posey

and HipSaver hip protectors, he stated, "... these are not definitive test data as exactly described in any forthcoming Standards as the position procedure has not been confirmed as the Posey pad appeared to be outside the area ... and could not be positioned in the rig." Professor Minns reiterated later that, "I hope that is clear and consequently cannot be referred to test data complying with the impending Standard until the positioning procedure has been clearly defined." The correspondence does not elaborate or provide any reasons as to why the pads could not be positioned for the test. Without such information, I find no basis to conclude with reasonable certainty that the statement made by the Posey ad, that the pads are "... positioned precisely over the hip bones ..." is true or false.

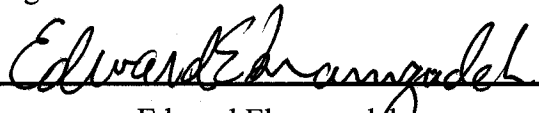
#### IV. POSSIBLE ADDITIONAL ANALYSIS AND INVESTIGATION

In support of my opinions, I may rely on visual aids and other demonstrative exhibits which may include, among other things, the attached Supplemental Exhibits, excerpts from deposition or trial testimony, documents and exhibits relied upon by other witnesses, additional information such as summaries of data from the materials I reviewed, or other types of materials.

I reserve the right to supplement this report in the event that additional information is provided to me. I may also rely on testimony given or to be given by other witnesses and on other reports and/or documents supplied to me in the future.

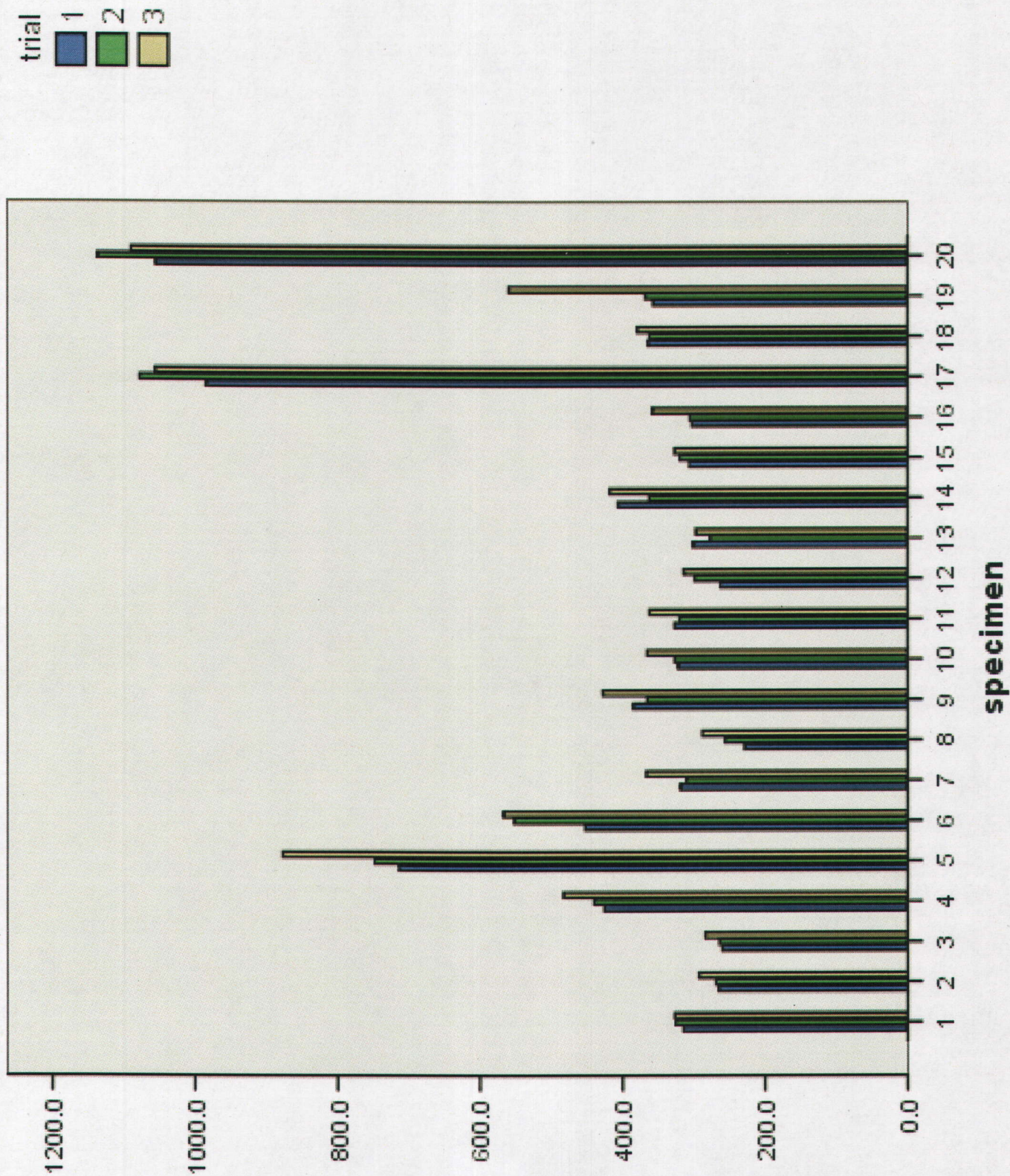
I reserve the right to perform additional investigations in this matter.

DATED: October 18, 2006

  
Edward Ebramzadeh

**SUPPLEMENTAL EXHIBIT "1"**





# EXHIBIT C

## PART 2



**SUPPLEMENTAL EXHIBIT "2"**

# Instructional Course Lectures

Volume 43 1994

Edited by  
Michael Schafer, MD  
Ryerson Professor and Chairman  
Department of Orthopaedic Surgery  
Northwestern University Medical School  
Chicago, Illinois

With 403 illustrations



American Academy  
of Orthopaedic Surgeons

## C H A P T E R 6 0

# Challenging the Validity of Conclusions Based on P-values Alone: A Critique of Contemporary Clinical Research Design and Methods

Edward Ebramzadeh, MS

Harry McKellop, PhD

Frederick Dorey, PhD

Augusto Sarmiento, MD

## Introduction

In the last few decades, clinical investigators have become increasingly aware that the use of statistical methods to assess the reliability of the outcome of a study forms an essential part of the scientific decision-making process. Unfortunately, among clinical researchers, there is a widespread misunderstanding of the meaning of the term, "statistical significance." Despite a consistent and emphatic outcry from statisticians and epidemiologists, the widely held misconceptions that  $P \leq 0.05$  is synonymous with statistically significant, and that statistically significant is synonymous with clinically significant or scientifically significant, have permeated the community, and are responsible for serious, ongoing damage to the scientific literature.

Several articles have been published in the medical, statistical, and epidemiologic literature in which the authors have attempted to clarify the meaning of P-values and statistical significance for the biomedical researcher.<sup>1,2</sup> In spite of these efforts, many investigators, as well as editors and reviewers of medical journals, continue to misinterpret the meaning of the P-value and the role it plays in the overall decision-making process.

This chapter seeks to convince the reader of two concepts. First, we will demonstrate that a P-value considered alone does not distinguish important from unimportant results and, second, that phrases such as "a statistically significant difference was observed" or "no significant difference was measured" would best be eliminated from the clinical and biomedical literature altogether. To accomplish our goal, we will need to discuss some issues involved in research design and data analysis.

Throughout the chapter, we will assume that the scatter in the measurements in our examples results primarily from the natural variation in the subject population from which the samples were drawn, rather than from error caused by inadequate precision of the measuring instruments, mistakes in calibrating machines or in logging the data, or other types of measurement error that might lead to bias. Therefore, the

P-value reflects the probability that the observed difference occurred by chance, that is, that we randomly selected subjects from the experimental and/or control groups that were not representative of the parent populations from which they were drawn.

## Shortcomings of P-values

### A Hypothetical Study

We will begin with a simple example to demonstrate one of the most common decision-making processes using statistical methods. A clinical investigator was interested in comparing the effectiveness of a newly developed drug to treat a specific fatal type of infectious disease. This investigator conducted a randomized, prospective study, comparing the outcome for two similar groups of patients, one treated with drug A (the experimental drug), and the other with drug B, a conventional drug that had been used with limited success over the last few years. Assume for the moment that drug A costs somewhat more, and that neither drug produces any significant side effects.

The investigator found that the rate of recovery from the infection was 30% higher in the group treated with the experimental drug A than in the group treated with the conventional drug B. However, a large variation in the grade and progression of the infection among patients and the relatively small number of patients in each group caused the corresponding P-value for the measured difference in recovery rates to be 0.11, well above the level of 0.05 that is often used to designate statistical significance.

### A Wrong Decision Based on the P-value

In the subsequent publication of the above study, the investigator reported simply that "There was no statistically significant difference in mortality between patients treated with drug A and those treated with drug B ( $P > 0.05$ )."

The investigator proceeded to recommend the continued use of the conventional drug B over the new drug A for three reasons: (1) because no significant difference was found, (2) because the

new drug costs more, and (3) because use of the conventional drug was standard, routine procedure. Based on this published information, several other institutions abandoned development of the experimental product.

The reader will surely appreciate that something is seriously wrong with the above application of the concept of statistical significance. Although the above form of decision-making is common practice, it represents a serious misuse of statistics that, nevertheless, is often encouraged or even required by the reviewers and editors of prominent medical journals. Clearly, if treatment with drug A has the potential to save the lives of more of the patients who have this disease, and there are no apparent side effects, then these results should not be dismissed in such an absolute way. Reporting simply that "no significant difference in mortality was noted between the two groups" would be a serious disservice to the medical and scientific community and, most importantly, to those patients whose lives might be saved by treatment with the new drug.

Unfortunately, this is precisely the type of interpretation, and terminology, that has come to permeate the clinical literature. For example, as a result of commonly held misconceptions regarding statistical significance, some journals have misguidedly come to require that the word "significance" be used *only* in relation to a *statistically* significant result, without regard to clinical or scientific significance,<sup>8</sup> and many journals allow the word significant only in association with P-values of less than 0.05.<sup>9</sup> Many have taken this type of confusion a step further by implying that the lower the P-value, the more important or valid the result. For example, an editorial in a prominent psychology journal<sup>10</sup> stated that, while results associated with a significance level of 0.05 may qualify for symposia and handouts, in order to qualify for archival journal publication, generally a P-value of 0.01 would be required.

For many investigators, having an associated P-value greater than 0.05 is considered to indicate that the observed difference is somehow an illusion, and that the statistical calculations fortunately save us from erroneously accepting the results as real. This is an absolutely wrong interpretation of the meaning of the P-value. In our example, as long as the result of the study is the only information that we have on the effect of drug A, the best estimate of the rate of recovery in a group of patients taking drug A is that it will be 30% higher than in the control group. Conversely, a P-value exceeding 0.05, or any other fixed value, does not mean that the rate of recovery will be the same for the two groups, or any difference other than 30%. As we shall discuss later on, in order to determine the stability and precision of this estimate, we need to calculate the confidence interval surrounding the 30% differ-

ence. While the consequences of such erroneous interpretations in the medical literature are typically more subtle than in our intentionally dramatic example, they nevertheless represent many misleading conclusions and serious loss of potentially important research knowledge.

### Objective Conclusions and Reporting of the Results

To further clarify this issue, the reader should imagine being among those clinicians who are desperately in search of a more effective treatment for the infectious disease in our example and should consider which of the four following reports represents the proper, logical deduction from the information at hand, and which is the most complete and accurate for reporting the results of the above study. (1) There was no significant difference in the recovery rate between the group taking drug A and the control group taking drug B ( $P > 0.05$ ). (2) The group taking drug A showed a statistically insignificant greater recovery rate ( $P > 0.05$ ). (3) The group taking drug A had a 30% greater rate of recovery, although this was not statistically significant ( $P > 0.05$ ). (4) The group taking drug A had a 30% greater recovery rate, with associated 95% confidence limits of -15% to 75% ( $P = 0.11$ ). Because this represents a potentially large and important clinical improvement, because there are no known deleterious side effects of taking drug A in this dosage, and because there was just over one chance in ten that the 30% difference occurred by accidental selection, we feel that it is justified to accept the results as true and to recommend the administration of drug A to patients with this disease while further studies are taking place.

The fourth statement clearly represents the most sound interpretation and conclusion. It also conveys the essence of the findings to the reader, and demonstrates how these findings are not overshadowed by the P-value in arriving at the decision to accept the results. Rather than an oversimplified statement of whether the results were "statistically significant" or not, the data, along with a range of likely values as well as the associated P-value, are presented, followed by an assessment of the scientific or clinical importance of the results.

In contrast, the first two statements provide virtually no useful information to the reader and, in fact, serve to hide the fact that a cure for the disease may be at hand. The third statement, while an improvement over the first two, still fails to address those issues that are more important than the P-value, such as what constitutes a scientifically or clinically important increase in the rate of recovery from the disease, what the potential benefits are of accepting the study's result as real, and what the potential losses are in rejecting it as false. Only after such issues are

addressed can one decide whether the certainty of the results, represented by the confidence interval and/or the P-value, is small enough to justify accepting or rejecting the results.

#### More on Decision-Making Based on P-values

There are historical reasons why results with P-values of 0.05 or less have mistakenly come to be taken as synonymous with statistically significant in so many studies today. The significance level of 0.05 was originally devised as an arbitrary convention by Sir Ronald Fisher,<sup>11</sup> who also laid the foundation for the formal process of testing for significance (hypothesis testing), which we will describe in some detail. Fisher, who has been called the greatest statistician who ever lived,<sup>12</sup> along with a number of most influential statisticians of our century, later denounced the widespread practice of considering any single significance level as always indicating statistical significance. Similarly, Kendall and Stuart,<sup>13</sup> writing in *The Advanced Theory of Statistics*, avoided the use of the word "significance" altogether, perhaps to prevent confusion between clinical and statistical significance.<sup>9</sup> In his book on clinical epidemiology, Feinstein<sup>9</sup> cites these and other eminent statisticians and mathematicians, including Yates,<sup>14</sup> Box,<sup>15</sup> and others, who have vigorously complained about how researchers commonly misuse statistical significance. Feinstein notes that in this regard, Box<sup>15</sup> referred to such researchers as being "overawed by what they do not understand" and who "mistakenly distrust their own common sense and adopt inappropriate (mathematical) procedures." As elaborated by Feinstein:<sup>16</sup>

"The [faulty] method of making statistical decisions about 'significance' creates one of the most devastating ironies in modern biological science...a clinician investigator will usually go to enormous efforts in mensuration. He will get special machines and elaborate technologic devices to supplant his old categorical statements with new measurements of 'continuous' dimensional data. After all this work in getting 'continuous' data, however, and after calculating all the statistical tests of the data, the investigator then makes the final decision about his results on the basis of a completely arbitrary pair of dichotomous categories. These categories, which are often called 'significant' and 'non-significant,' are usually demarcated by a P-value of either 0.05 or 0.01, chosen according to the capricious dictates of the statistician, the editor, the reviewer, or the granting agency. If the level demanded for 'significant' is 0.05 or lower, and the P-value that emerges is 0.06, the investigator may be ready to discard a well-designed, excellently conducted, thoughtfully analyzed, and scientifically important experiment, because it failed to cross the Procrustean boundary demanded for statistical approbation."

Elsewhere, Feinstein<sup>9</sup> writes:

"Because clinicians have a long tradition of avoiding any specifications for quantitative judgements....there are no medical standards for quantitative significance....The absence of standards of quantitative significance...has been a boon for investigators, pharmaceutical companies, editors, and regulatory agencies.... It has enabled the accomplishments....to be appraised exclusively according to standards of....[statistical]....significance....If P is < 0.05, the investigator is granted approval for the claim that the observed distinction is significant....Because the desideratum of P < 0.05 can always be obtained [by studying a larger sample], no matter how trivial the quantitative importance of the observed distinction, investigators who were wise enough (or fiscally supported enough) to study large groups have been able to achieve significance and to gain editorial or regulatory approval for claiming a significant action for agents that had minor importance in science or in clinical therapy. Conversely, however, an investigator who has found an agent of major quantitative importance may have his paper rejected for publication or his drug disapproved for marketing because one or two of patients dropped out of the study, thereby raising the P value above 0.05 to the disastrous height of 0.06. This policy continues to be applied....even though [it] has been disclaimed or condemned by Fisher himself and other prominent leaders in....statistics."

#### Bias in the Literature

In addition to preventing publication of potentially important, clinically significant research, these misguided policies have created a systematic bias in the literature. This occurs because only those papers that have shown a particular difference at a "statistically significant" level (commonly  $P \leq 0.05$ ) are published, even though a larger number of studies may have used the same method and measured the same type of difference, but obtained a lower level of statistical significance, indicated by a larger P-value. Thus, when one conducts a survey mathematically combining the results of several studies (technically called a meta-analysis),<sup>17</sup> the prior selection of those studies with low P-values for publication creates a false impression of the overall level of certainty associated with the observed difference.

Conversely, if several investigators, working independently and using similar methods on similar samples, each observe a particular difference, but with P-values that, individually, would not justify accepting the results, the combined results (obtained by a meta-analysis) may provide an acceptable level of certainty. The reason for this combined difference may be clarified by referring to the fundamental principles of probability.<sup>17</sup> Each time that a particular apparent



result is obtained in independent investigations, the combined probability of the difference having occurred by chance selection is decreased. It is through this process that the scientific method is used to increase progressively our confidence in an observed difference when that difference represents a true effect, particularly when the number of samples included in each individual study is small and the inherent scatter is large. This illustrates yet another important reason for not adopting a significance level of 0.05, or any other specific value, as a criterion for publishing scientific results.

### **Scientific Deduction Versus Decision Making**

In order to understand how we generally come up with decisions following observation of a particular difference, we need to clarify our thinking process. Consider an everyday occurrence in clinical practice; to select between two or more methods of treatment for a particular case. Suppose that you are faced with making such a decision, and that you have just read the results of a study, such as our initial example above, that appears to show an advantage with one of two treatments. Rather than a single decision, there are now two distinct considerations at hand. First, you need to assess what can be learned from the information obtained with the sample, and whether you believe the results of the study to be an accurate representation of the true difference in effectiveness of the treatments in the general population. To assist in making this first decision, you may use statistical methods such as a P-value, confidence intervals, or other measures, as we will discuss later on. Once you have made the decision on whether or not you accept the results of the study, the second decision is what course of action you should take. This decision is made partly based on whether or not you believed the results to be representative of the differences, but also on risk-benefit assessments and other practical considerations. As Edwards and associates<sup>18</sup> stated, "Sometimes the.... definition of [statistical] testing is expressed as a decision to act as though one of the two hypotheses were believed, and that has apparently led to some confusion. What action is wise of course depends in part on what is at stake. You would not take the plane if you believed it would crash, and would not buy flight insurance if you believed it would not." Choosing to take a course of action other than what the results of the study suggested is not necessarily equivalent to believing that the results of the study are false. In short, we need to distinguish between our judgment on the data on the one hand, and our decision regarding an action on the other.

To clarify this point, consider an example of a mountaineer descending a glacier and faced with the

choice of either jumping across a narrow but very deep crevasse, or of walking several tiring miles around the crevasse to arrive at his camp. The mountaineer has no rope, so failing to jump completely across the crevasse will result in his death. However, he has practiced such jumps wearing a rope in the past and knows that his mean jumping distance is twice the width of the crevasse, with an associated P-value of 0.001. Thus, while he is almost certain that he can jump across the crevasse (he believes his data to be a true representation of his jumping skill in general), because there is a small probability that his data was obtained by chance, combined with the severe penalty of death if it is wrong, he chooses to walk around the crevasse.

Now consider a second mountaineer faced with the same decision. He is not quite as consistent a jumper as the first, such that, although his mean jumping distance is also twice the width of the crevasse, the associated P-value is 0.1. Although he at first also chooses to walk around the crevasse, a nearby television crew offers him one million dollars to make the jump on camera. With this in mind, the second mountaineer now chooses to attempt the jump.

Comparable situations arise in medical science when an apparent experimental difference is demonstrated with a very low P-value that, if correct, would cause a major change in current theory or organizational policies. That is, if the results are accepted, then it will be necessary to rewrite textbooks, to redesign experimental equipment to take the effect into account, and to reallocate millions of dollars of research funds in this area. Because this would be a major disaster if it was subsequently learned that the apparent difference occurred by chance, the scientific community may justifiably choose to reject the results, despite a P-value of, say, 0.001, until several independent investigators have obtained similar results with comparably low P-values. Once again, however, rejecting such results is not the same as believing them to be false, in spite of whether the associated P-value is 0.001 or 0.1. Conversely, if an apparent measured difference is consistent with present theory (or, if there are no contradictory results or theory) and the penalties for a wrong decision are minor, it may be appropriate to accept results that have potential scientific or clinical importance, despite a relatively high P-value.

In the above examples, the mountaineers used the P-value to assess the risk of being wrong in concluding that they could jump over the crevasse. For the first mountaineer, there was a one in a thousand chance of being wrong in concluding that he could jump successfully; for the second mountaineer this chance was one in ten. Obviously, in order to arrive at a decision, it was more useful for the mountaineers to know the actual P-value, which represents the risk that they are taking, rather than only if the P-value was less or more than

0.05, or some other specific value, which is all they would know if they were only told the result was "significant" or "not significant."

### **The Controversy Over Hypothesis Testing**

Hypothesis testing was adopted as an indispensable tool for industrial applications and quality control. For instance, in factory assembly lines, a decision needed to be made regarding the quality control based on a small sample of the day's lot. The choices were distinct: accept the day's lot or reject it. Therefore, a standard, routine, simple and relatively thought-free<sup>19</sup> way of making this decision needed to be developed. The answer was to measure a critical value with the sample, and base the decision on whether the difference between the measurement obtained with the sample and a control value was statistically significant or not. This facilitated making many decisions that would, in the long run, produce a controlled amount of variation in the outcome.

In contrast to industrial applications, in the scientific world, measurements are performed primarily to further our knowledge and understanding of how the world works.<sup>19</sup> Although we certainly use this knowledge in our daily decision making, we must distinguish between the process of making judgments about the data and decision making. The distinction is subtle but important, because it affects the way we interpret our observations and decide on the consequent course of action.<sup>19,20</sup> To clarify the issue, consider once more the example of the mountaineers who were faced with either jumping the crevasse or walking a long, tiring distance. Each one made a decision which was partly influenced by the knowledge gathered from previous experience but was primarily based on practical issues that had little to do with statistical significance. Because we are human, our decisions are to a large degree governed by our desire to obtain the benefits and by our fear of the penalties, and, therefore, they can at times be far from objective.

These considerations have led to two schools of thought regarding scientific research design and data analysis. Some statisticians still believe that, as in the factory setting described above, hypothesis testing is a useful tool for research.<sup>21</sup> That is, if we set up the study appropriately, we can and should determine our course of action, such as which treatment to use, based on the scientific and statistical significance of the data. These statisticians still recommend hypothesis testing, arguing that a standard method of decision making is required. This point of view, while apparently held by very few statisticians today, unfortunately still remains the dominating viewpoint among the majority of clinical researchers.

The more popular perception among statisticians, one that is not yet fully recognized by many

researchers, is that scientific research design and data analysis should not necessarily include a strict decision on accepting or rejecting a hypothesis. Statisticians have long argued that the process of data evaluation by hypothesis testing is so erratic and subjective and can be so misleading and deceptive that it would best be abandoned entirely in scientific applications.<sup>1,2,22</sup> Rather, the goal in scientific research should be to extract as much useful information as possible from the sample data to assess meaningful judgment on the general population.

Even statisticians who encourage hypothesis testing agree that it has been subject to misinterpretation, and therefore may be misleading, in at least some situations. For example, Fleiss, who wrote in defense of the use of significance tests,<sup>21</sup> nevertheless started his article with the following: "There is no doubt that significance tests have been abused.... statistically significant associations have, in error, been automatically equated to substantively important ones, and statistically nonsignificant associations or differences have, in error, been automatically equated to ones that are zero." Unfortunately, alternative solutions, such as data evaluation using confidence intervals, have also been subject to misuse, as will be described. Regardless of whether one decides to conduct a formal hypothesis test or not, one must understand the underlying principles involved in risk-benefit assessment, which is the only way to select a proper sample size for a given study.

In the next sections, we will discuss the issues involved in risk-benefit assessments and in estimating the precision of our measurements. For thoroughness, we will describe the issues involved with and without a formal hypothesis test. It must be emphasized that mathematically the procedures are very similar, whether or not a formal hypothesis test is performed; the difference lies mostly in the interpretation of the results and in the deriving of conclusions.

## **Assessing the Uncertainty in Our Data**

### **Generalizing Conclusions From Experimental Study Groups**

Consider our initial hypothetical study again, which compared the effectiveness of two drugs for the treatment of a particular disease. This study evaluated the difference in the recovery rate with the two drugs in an experimental sample, but its final goal was to estimate the difference in the effectiveness of the drugs in the general population. In other words, we would like to use the difference obtained with our selected group of patients to predict the difference that would be obtained if we used the same treatment in all patients with the disease, now and in the future.

Because we found a 30% difference in recovery rate between patients treated with the two drugs, we esti-

mate that the difference in the general population will also be 30%. This remains the best estimate of the outcome in the general population, regardless of the associated P-value. Similarly, in the example of the mountaineers, the best estimate of how far either one would jump remained twice the width of the crevasse, and having a higher or lower P-value did not change what was the best estimate of the outcome.

While the sample provides us with the best estimate of the difference in the larger population, this is still only an estimate. The investigator in the drug study should not expect to see an improvement of exactly 30% with the new drug in the general population, and neither mountaineer should expect to jump exactly twice the width of the crevasse in subsequent attempts. The fluctuation of distance for various attempts might be extremely small or very large. While a large P-value suggests (but does not show unequivocally) that there is perhaps more of this type of fluctuation, it provides no indication of the magnitude of this fluctuation. We use confidence intervals to determine the possible magnitude of fluctuation of treatment effects.

### Confidence in Our Estimation

Just how precise is our estimate of a difference (in ratios or means) obtained from an experimental sample? Surely, if a difference of 30% is obtained between two groups of 1,000 patients each, it is more reliable than the same difference if it were obtained with ten patients in each group. The P-value provides us with the probability that our difference was obtained by accident, but does not provide us with a measure of variation for the obtained difference.

The confidence interval can be used to estimate a range of numbers within which the true difference (the difference in the general population) lies. In order to calculate the confidence interval, we must first decide what level of certainty we require. Suppose, in our drug test example, that we calculate the 95% confidence intervals for the difference in recovery from infection, and we obtain a range of between 15% and 45%. This would mean that we can be 95% certain that the true difference in recovery rate in the general population will be between 15% and 45%. If a lower level of certainty is acceptable, such as 90%, then one could calculate the 90% confidence interval. The lower the level of certainty that is required, the smaller the confidence interval.

Confidence intervals, which give a range of likely values for the true difference, provide more insight into the nature of the data than do P-values.<sup>2,4,6,22</sup> Statisticians have encouraged researchers to use confidence intervals for comparisons in addition to or instead of assessing statistical significance using P-values. Whether one uses P-values or any other statistical measure, confidence intervals can always provide more

useful information. However, one must bear in mind that each confidence interval is based on a fixed probability level. Making a comparison using 95% confidence intervals will point to the same conclusion as if the comparison were made using  $P \leq 0.05$  to indicate statistical significance. The investigator must consider the difficult yet important question of what confidence level to use for a given study.

### Estimating the Appropriate Sample Size

As demonstrated with our examples, the decision-making process in a scientific study is much more complex than simply inspecting whether the P-value is smaller or greater than some reference value such as 0.05. A firm understanding of each of the steps in this process is fundamental to arriving at sound conclusions. Most investigators are advised that they must establish a clear, unambiguous statement of a hypothesis. However, too few researchers pay sufficient attention to other vital questions that need to be answered in advance of the study, regardless of whether they want to establish a formal hypothesis test or not. Specifically, before each study, an investigator must answer the following questions:

(1) What is the smallest observed difference (for example, difference in survival rate for patients taking drug A or B) that one could consider to be of scientific, clinical, or practical importance? This value is referred to as  $\delta$  (delta). This value may be difficult to assess, but if one cannot state before the study begins how much of a measured difference one will consider to be clinically significant, this is not likely to be a simpler matter after the results are obtained, and one should question whether the study should be conducted in the first place.

(2) What will be the scientific or clinical benefits of correctly accepting the observed difference as representative of the general population, and what will be the penalties for incorrectly accepting the difference as real?

(3) What will be the benefits of correctly rejecting an observed difference as having occurred by chance? What will be the penalties of incorrectly rejecting the observed difference as having occurred by chance?

(4) With what certainty does one need to know the difference, and what magnitude of error in the difference would be acceptable?

The required certainty is called the confidence level. For example, we may need to know the increase in serum glucose in patients following a treatment with a particular type of oral hypoglycemic. If we select a 90% confidence level and an error of 25 mg/dL,<sup>2</sup> we are stating that we would like to be in error by no more than 25 mg/dL<sup>2</sup>, with an associated assurance of 90%. Mathematically, selection of the error is equivalent to



the selection of the smallest clinically or scientifically significant difference,  $\delta$ , which was the answer to the first question. Stated differently, if we select an error of 25 mg/dL<sup>2</sup>, we would like to know the serum glucose to within 25 mg/dL<sup>2</sup>, which in turn would mean that a difference of smaller than 25 mg/dL<sup>2</sup> would not be clinically significant. Therefore,  $\delta$  and the error are practically the same parameter, and these cannot be selected independently, though they have different interpretations.

(5) If a formal hypothesis test is to be performed, one must answer the following. Given the above assessments, for a given magnitude of observed difference, what is the maximum probability that the observed difference occurred by chance selection that one is willing to allow and still accept that the difference is representative of the general population?

The answer to this fifth question is called the significance level,  $\alpha$  (alpha), and is the number to which we later compare our obtained P-value. As discussed earlier, based on traditional but misleading terminology, if the obtained P-value is smaller than alpha, then the measured difference is referred to as statistically significant, whereas, if the P-value exceeds alpha, the results are called not statistically significant. Also based on tradition, the alpha level is widely, but often inappropriately, chosen as 0.05, implying that the investigators are willing to accept no greater than a one-in-twenty probability that the observed difference occurred by chance alone. However, for a given study, the appropriate level of certainty may be well above or well below 0.05.

Mathematically, selecting a significance level  $\alpha$  of 0.05 is equivalent to selecting a 95% confidence level. Therefore,  $\alpha$  and the confidence level cannot be selected independently, although they have somewhat different interpretations.

The considerations in the assessment of the above risk may include whether the apparent results contradict established scientific principles or the results of earlier studies, whether there are dangerous side effects of a drug or treatment, the cost of the treatment, and others. Thus, depending on the consequences of accepting results that may have occurred by chance, the investigators may choose to require a very small significance level, such as 0.01, or they may be comfortable with a higher probability that the apparent difference occurred by chance such as 0.25, or one chance in four. Equivalent statements to these would be to select a high confidence level, such as 99%, or a low confidence level, such as 75%, respectively.

Unless the significance level ( $\alpha$ ) is selected in advance, and is based on careful assessment of risk-to-benefit ratios, there is little point in dividing results into categories of statistically significant and not statistically significant after the study has been conducted. Particularly in such cases, it is most appropriate to

report the difference obtained along with the associated confidence interval and P-value, rather than simply to report whether the result was statistically significant or not. By the same token, it is inappropriate for the editors of a scientific journal to designate a single significance level, such as 0.05, or a single confidence level, such as 95%, as a universal designation of statistically significant results, because, in effect, the implementation of this significance level may be occurring well after the results are obtained from the study.

(6) Given the above assessments, for a given magnitude of observed difference, what is the maximum probability that one is willing to allow that one will fail to detect a clinically significant difference when in fact there is one? The answer to this question is called  $\beta$  (beta), the probability of a type II error. The importance of type II errors in clinical research has been emphasized by many authors.<sup>3,7,23</sup> The probability of not making a type II error is called the power of the test. Suppose, in our example of the comparison of the effectiveness of two drugs, that the power of the test was 83%. This would mean that there is an 83% probability that the 30% difference that we obtained with the sample was truly representative of the difference in the general population.

When one selects the significance level ( $\alpha$ ), the tendency may be to select the smallest value that the sample study allows. An equivalent practice is to select the confidence level as high as possible. However, the investigator must realize that, for a given sample size, selecting a smaller significance level means increasing the probability of type II error. In other words, the smaller the chances one chooses to take in erroneously concluding that there is a difference, the larger the chances that one would fail to detect a difference when in fact there is one. The investigator, then, must wisely consider the seriousness of each type of error, and select  $\alpha$  and  $\beta$  accordingly. In addition to answering the above questions, the investigator must either measure or somehow estimate the variance(s) in the groups being considered.

It should be obvious by now that the selection of the smallest outcome that would be considered clinically important (the first question) and the assessment of the risk-to-benefit ratios that lead to the selection of the required levels of certainty ( $\alpha$  and  $\beta$ ) are not arbitrary, although they may be subjective. These decisions must be made taking into account all of the available relevant information, including fundamental theory, existing literature on the subject, prior experience of the investigators, and, above all, common sense.

The answers to all six questions must be provided by the investigators, with the collaboration of a qualified statistician, who may then apply the appropriate mathematical formulas for each case in order to select the minimum sample size (number of patients) that would

be required to detect the magnitude of the difference in question at the required level of statistical certainty. Note that the sample size can be selected to limit the error (based on confidence intervals), or to obtain the desired power.

Selection of the appropriate sample size is perhaps the most important part of the preparation for a study. If the sample size is too small, even very large measured differences may have associated P-values greater than the alpha level, such that the result would be considered statistically insignificant. Conversely, with an excessively large sample size, even small differences with little or no scientific or clinical importance may, nevertheless, be measured with a very high level of statistical significance. In addition to being misleading, the latter event may represent a major waste of scarce research funds.

### Evaluation of Data Using Confidence Intervals

Statisticians have been advocating the use of confidence intervals instead of (or in addition to) P-values,<sup>2,4,6,22</sup> because they can be more informative, illustrative, and less misleading than P-values. As a first step, given the confidence level that was determined before the study, such as 90%, we can calculate the confidence interval of the difference that was obtained with our sample. This confidence interval provides us with a range of likely values for the difference. Technically, we could use the confidence interval to evaluate the statistical significance of the difference. That is, if the confidence interval of the difference includes zero, the obtained difference is not statistically significant; and on the other hand, if the confidence interval is on one side of zero only, the result can be considered statistically significant. Mathematically, if we had selected a 90% confidence interval, this comparison would be the equivalent of selecting a significance level of 0.1. However, it would be a waste of resources if we only used confidence intervals to assess statistical significance. As indicated, the range of numbers provides us with the proper perspective of the likely values for the difference. Thus, we should consider the entire range of values of the confidence interval, rather than simply the end points. In contrast, considering the P-value alone is similar to looking at only one end of the confidence interval.

Because the confidence interval includes a range of likely values for the difference, it can be used to evaluate clinical or scientific significance along with statistical significance. That is, if the lower end of the confidence interval of the obtained difference is larger than the magnitude of smallest difference that we consider clinically significant, then the obtained difference is clinically significant; and on the other hand, if the confidence interval of the obtained difference is entirely

below this value, then the result is clinically insignificant. If the minimum value that we consider to be clinically significant lies somewhere within the confidence interval, we require further study. In addition, the confidence interval of the difference provides us with some indication of the power of the test. However, as with the assessment of statistical significance, the entire range of values that the confidence interval represents must be considered, and not simply the end points.

Unfortunately, in the literature, analysis with confidence intervals is plagued with problems similar to those with P-values.<sup>19,24</sup> Most authors use only 95% confidence intervals, rather than 75%, 80%, 90% or any other value, without regard to the risk-benefit considerations that we have already discussed. In order for the conclusion to be valid, the level of confidence must be selected, on the basis of risk-benefit assessment as described above, before conducting the study. Furthermore, rather than paying attention to the approximate range of values that the confidence interval provides, and where the larger portion of the ranges lie, many simply inspect whether zero is inside or outside of the confidence interval of the difference. This oversimplified practice, as described, may lead to many misleading conclusions. Technically, this would be no different from concluding that, say, a P-value of 0.051 indicates an insignificant result.

### Evaluation of Data for Hypothesis Testing

#### Evaluation of Clinical Significance

As mentioned in the previous section, the first step in the planning of a study should be to select  $\delta$ , which is the smallest difference that would be of medical, practical, or clinical importance. After the study has been conducted, the first step in the evaluation of the results should be to compare the obtained difference to  $\delta$ , in order to assess whether the observed difference should be considered of scientific or clinical importance. Unless confidence intervals are used, this may be the only step in which one considers the consequences and thinks directly about the magnitude of the obtained differences.

If the observed difference is larger than  $\delta$ , the result is considered clinically or scientifically significant, and we may proceed with the calculations of statistical significance in order to assess the reliability of our assessment. If, on the other hand, the observed difference is smaller than  $\delta$ , then the result is likely clinically insignificant. In this case, the data suggest that the groups being compared are similar, at least for practical purposes. Again, we need to assess the statistical reliability of this assessment, as explained below.

Although the evaluation of the clinical significance is the most important step in the interpretation of the

results of a study, this is commonly ignored entirely in studies published in the medical literature. For example, if one has measured the blood cholesterol level in patients following a certain diet, is the magnitude of the measured decrease or increase in the level of cholesterol large enough to make a considerable difference in the risk of heart disease? If one is comparing the longevity of a new design of total joint replacement with that of conventional designs, is an increase (or decrease) in longevity of a month or two important enough to justify recommending the new design? If not, would an increase of one year be sufficient? How about two years? The only way to objectively answer these types of questions is to decide on and document what would constitute a clinically significant outcome before starting the study. Otherwise, once the study is done, the investigator may be compelled to exaggerate the importance of small (though perhaps statistically significant) differences in order to justify the time and money spent on the study.

#### Considerations in Accepting That a Difference Exists

If the magnitude of a difference obtained in a study is determined to be of scientific or clinical significance, then one should evaluate the reliability of this difference. In other words, we should evaluate the risk of making the assessment that the difference with our observed sample is representative of the true difference in the general population. This is commonly done using P-values. A difference that is clinically significant should never be dismissed, regardless of the P-value.

Supposing that the difference is clinically significant, let us consider the possible outcomes. If the P-value is smaller than the predetermined significance level, then the result is considered statistically significant. Most investigators manage this type of result quite well, particularly if they are using the traditional significance level of 0.05. On the other hand, if the P-value is larger than 0.05 (say, 0.08), there is much confusion among investigators regarding how to interpret and report the results, or even what kind of conclusions to make from the finding. Some regard the reporting of differences associated with P-values of larger than 0.05, however large and important, as "controversial," "liberal," "weak," or "lowering the (established) statistical standard." However, the only objective way of reporting such results is to report the obtained difference along with the associated P-value. There are strong reasons to thoroughly consider the likelihood that such results are true representations of the general populations from which they were drawn.

When the obtained difference is clinically or practically significant, and the associated P-value is too large to indicate statistical significance, the most likely reason is that the sample size was too small. In other words, other investigators may repeat the study with

larger samples and eventually obtain statistical significance. On the other hand, if the result was obtained only by chance, that is, if the difference does not exist in reality, that, too, will be determined by other, perhaps larger, studies. The important point is that not reporting obtained differences because the P-value was larger than what is considered statistically significant may hinder the progress of science by hiding potentially important results from other investigators.

#### Establishing That Groups are Similar

When the obtained difference is smaller than what is considered clinically significant, it suggests that the groups being compared are samples from functionally similar populations. To validate this, it is common for investigators to erroneously use P-values alone. In doing so, investigators run a potentially high risk of arriving at the wrong conclusion that no difference exists in the general populations, when in fact there is a large difference (a type II error). This erroneous type of deduction is quite common among clinical investigators, despite clear warnings against it in the literature.<sup>7,23</sup>

As an example, Freiman and associates<sup>23</sup> analyzed the results from 71 clinical trials, published in well-known international medical journals, all of which had reported "no significant difference,  $P > 0.05$ " between the study and the control groups. Freiman and associates demonstrated that 67 of the 71 trials had a greater than 10% chance of failing to detect a true 25% therapeutic improvement with the study group compared to the control group, and that 50 of the trials had a greater than 10% chance of failing to detect a true 50% improvement. When these authors calculated the 90% confidence intervals for the true improvement in each trial, 57 of the treatments had a potential for 25% improvement over the control, and 34% had a potential for 50% improvement. Despite these potential improvements, only one of the studies had mentioned that both  $\alpha$  and  $\beta$  were considered prior to the start of the trial. In short, Freiman and associates concluded that it is quite likely that a large number of the studies included in their survey would reveal a true benefit with the treatment they were studying, had they continued the clinical trial longer. This study vividly demonstrated the potential risks in making conclusions based on P-values alone. Such incorrect analyses account for substantial loss of scientifically and clinically important information.

To illustrate the proper analysis when the results indicate similarity rather than difference, and when a hypothesis test is necessary, consider the following. A study published in 1981 compared the risk of death as a result of lung cancer among wives of over 175,000 smoking and nonsmoking men.<sup>24</sup> In this 12-year study, it was found that, compared to the wives of nonsmok-



ers, there was 10% higher mortality from lung cancer among wives of heavy smokers ( $\geq 20$  cigarettes per day), but the difference was not statistically significant. (Diamond and Forrester<sup>26</sup> have reanalyzed the data from this and several other studies using posterior probability based on Bayes' theorem, and have demonstrated some additional shortcomings of P-values beyond those discussed here. The discussion of posterior probability is beyond the scope of this article, as is the discussion of the literature on passive smoking. The purpose of citing Garfinkel's study is to demonstrate the underlying concepts of clinical significance, delta and power.) Suppose that, at the outset, the investigators had decided that any increase (or decrease) greater than 5% in the risk of lung cancer should be considered clinically significant. In other words, they set  $\delta$  to be 5%. In this case, the difference of 10% increase in lung cancer obtained in the sample would be clinically significant. However, they reported the result as not statistically significant, that is, they obtained a P-value that was larger than the significance level they had selected. When the result is clinically significant but not statistically significant, as discussed in the previous section, further study is required to determine whether the difference was obtained by chance, or whether it can be attributed to the general population. In any case, the difference cannot and should not be ignored based on the fact that it was not statistically significant. The most likely reason for this type of result (clinically significant but not statistically significant) is having a sample size that is too small. In such cases, the value of  $\beta$  becomes critical in determining what conclusion can be made from the study.

On the other hand, suppose that the investigators had decided at the outset that a difference in mortality from lung cancer has to be at least 20% before it can be considered clinically significant, that is,  $\delta = 20\%$ . (In reality, such a decision should have a reasonable basis, the purpose here is to demonstrate a different condition of the evidence using the same example). In this case, a 10% difference is not large enough to be considered clinically or practically significant. Therefore, because we have not observed a sufficiently large difference with the wives of smokers and non-smokers that we happened to study, the data suggests the conclusion that the wives of smokers and nonsmokers are at similar or the same risk of lung cancer in the general population. The result in this case is neither clinically nor statistically significant. Again, we must consider the likelihood that this result was obtained because we had too few samples. We thus turn to other statistics to arrive at a conclusion.

What we need to evaluate in this case is the probability that a larger difference exists between wives of smokers and nonsmokers, which we failed to detect by studying this particular sample. This type of error, as

mentioned earlier, is called a type II error, and the chances we are willing to take in making such an error is called  $\beta$  (beta). As indicated, we must decide on  $\beta$  before beginning the study. However, unlike  $\alpha$ , which can and should be set exactly before the study,  $\beta$  can only be approximated in advance. When we have the results, we can calculate  $\beta$ . In practice, we more commonly deal with the power of the statistical comparison, which is the probability of not making such an error, and therefore correctly concluding that there is no difference. For instance, if  $\beta = 15\%$ , then the power is  $100\% - 15\% = 85\%$ .

### Hypothesis Testing Versus Confidence Intervals

We have described two methods of evaluating the data observed in a study and making conclusions about the general population. As indicated, the mathematical calculations are similar. However, as demonstrated, confidence intervals offer a straightforward, simple way of evaluating data, whereas analysis of data for hypothesis testing can become quite complex. One must recognize that even if we do not conduct a hypothesis test, and even if we only use confidence intervals to describe our data, we still need to select  $\alpha$ ,  $\beta$ , and  $\delta$  before the study in order to select the proper sample size.

Comparing two analyses of the same problem, one with 95% confidence intervals of the obtained difference and the other with the P-value compared to 0.05, the conclusion will technically be the same. If the difference is marginally approaching significance, the confidence interval of the difference will barely cross zero, indicating that most of the likely values of the difference will be larger than zero. In contrast, for the same analysis,  $P = 0.053$  does not convey a range of values for the difference, but rather just one probability value. Therefore, confidence intervals are more informative and less misleading, as long as they are presented in their entirety, and not the conclusion of significant or not significant, which will essentially be the same no matter which method is used.

### Guidelines for Proper Description and Interpretation of Results

It would be naive to imagine that any single solution would solve the problems of misuse of statistical analysis within a short period of time. However, the authors feel that, as indicated, hypothesis testing inevitably leads to incomplete description of the data. In order to encourage complete reporting of the results, and to help minimize some of the problems we have outlined in this chapter, we propose that the orthopaedic community adopt the following guidelines for evaluating and publishing the results of an investigation.

First and foremost, the terms statistically significant and not statistically significant should be avoided entirely in the literature. Editors should not allow authors to describe the results of comparisons simply as being either significant or not significant. Rather, the investigators should report the data obtained by the measurements (tabulated, plotted along with standard deviations, etc.) along with the actual values of the associated confidence intervals and P-values. In discussing the observed differences, the authors should address the following issues: (1) What magnitude of difference was measured? (2) How does the magnitude of the observed difference compare to one that constitutes a scientific or clinically important difference? (3) If the measured difference is large enough to be scientifically or clinically important, what was the associated level of certainty for the measurement? In particular, what is the corresponding confidence interval of the difference obtained? (4) What will the benefits be if the investigator is correct in accepting the results? What are the penalties if the investigator is wrong? (5) Considering the above assessments, and based on the results, what conclusion can be made regarding the general population? In addition to the above, it is always beneficial to indicate the associated P-value, but report the P-value itself rather than comparing it to a significance level.

### Conclusions

Sound research practice requires a great amount of preparation. While some researchers mistakenly consider statistical analysis as a task that needs to be performed after the study is concluded to see if the result was significant or not, the correct way to design a study is to appraise the risk-to-benefit ratio and estimate the required sample size before the study is commenced. In particular, the smallest difference that would be considered scientifically, practically, or clinically significant should be specified in advance of the study.

In reaching a conclusion regarding the outcome of a study, the researcher must take into account all of the facts, figures, and evidence available and, above all, must apply common sense. For example, if the results directly contradict long-established scientific principles, then they may be viewed with justified skepticism, even though very strong statistical significance is obtained, at least until the same results have been obtained by further research or by other independent investigators. It should be kept in mind that, even at  $P = 0.05$ , one out of every 20 observations will have occurred by chance. On the other hand, if an observed difference is consistent with previously established scientific principles, follows logical deduction, and/or has been demonstrated in similar studies by others, it may be reasonable to accept the difference as real,

regardless of a relatively large P-value for an individual study. Finally, sound research requires a fundamental understanding of the statistical methods that are used. Most importantly, whether the investigator chooses to accept or reject the results, all of the results of the study should be reported along with the associated confidence intervals and P-values, regardless of whether these meet one's criterion of statistically significant.

### Acknowledgements

This study was supported in part by the Los Angeles Orthopaedic Foundation. The authors wish to thank Patricia Normand for her contribution.

### References

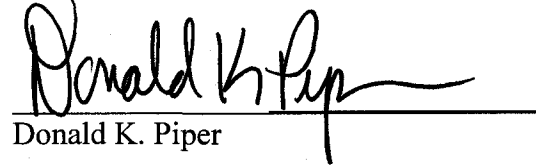
1. Walker AM: Reporting the results of epidemiologic studies. *Am J Public Health* 1986;76:556-558.
2. Rothman KJ: Editorial: A show of confidence. *N Engl J Med* 1978;299:1362-1363.
3. Rennie D: Editorial: Vive la difference (P less than 0.05). *N Engl J Med* 1978;299:828-829.
4. Thompson WD: Statistical criteria in the interpretation of epidemiologic data. *Am J Public Health* 1987;77:191-194.
5. Altman DG: Statistics and ethics in medical research: VII. Interpreting results. *Br Med J* 1980;281:1612-1614.
6. Dorey F, Anstutz H, Nasser S: The need for confidence intervals when presenting orthopaedic data. *J Bone Joint Surg*, in press.
7. Lieber RL: Statistical significance and statistical power in hypothesis testing. *J Orthop Res* 1990;8:304-309.
8. Ingelfinger FC: Editorial: Significance of significant. *N Engl J Med* 1968;278:1232-1233.
9. Feinstein AR: *Clinical Epidemiology: The Architecture of Clinical Research*. Philadelphia, PA, WB Saunders, 1985.
10. Melton AW: Editorial. *J Exp Psychol* 1962;64:553-557.
11. Fisher RA: *Statistical Methods and Scientific Inference*, ed 2. Edinburgh, Oliver and Boyd, 1959.
12. Brown BW: Statistics, scientific method, and smoking, in Tanur JM, Mosteller F, Kruskal WH, et al (eds): *Statistics: A Guide to the Unknown*, ed 2. San Francisco, Holden Day, 1978, p 66.
13. Kendall MG, Stuart A: *The Advanced Theory of Statistics*, ed 4. New York, Hafner Press, 1977, vol 1, 1979, vol 2, 1983, vol 3.
14. Yates F: Theory and practice in statistics. *J R Statist Soc* 1968;131:463-475.
15. Box GEP: Science and statistics. *J Am Statist Assoc* 1976;71:791-799.
16. Feinstein AR (ed): *Clinical Biostatistics*. St. Louis, MO, CV Mosby, 1977, chap 5, pp 54-70.
17. Rosenthal R: Combining results of independent studies. *Psychol Bull* 1978;85:185-193.
18. Edwards W, Lindman H, Savage LJ: Bayesian statistical inference for psychological research. *Psychol Rev* 1963;70:193-242.
19. Poole C: Beyond the confidence interval. *Am J Public Health* 1987;77:195-199.
20. Cox DR, Hinkley DV: *Theoretical Statistics*. London, Chapman and Hall, 1974.
21. Fleiss JL: Significance tests have a role in epidemiologic research: Reactions to AM Walker. *Am J Public Health* 1986;76:559-560.
22. Gardner MJ, Altman DG: Confidence intervals rather than P values: Estimation rather than hypothesis testing. *BMJ* 1986;292:746-750.

23. Freiman JA, Chalmers TC, Smith H Jr, et al: The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial: Survey of 71 "negative" trials. *N Engl J Med* 1978;299:690-694.
24. Poole C: Confidence intervals exclude nothing. *Am J Public Health* 1987;77:492-493.
25. Garfinkel L: Time trends in lung cancer mortality among nonsmokers and a note on passive smoking. *J Natl Cancer Inst* 1981;66:1061-1066.
26. Diamond GA, Forrester JS: Clinical trials and statistical verdicts: Probable grounds for appeal. *Ann Intern Med* 1983;98:385-394.

**CERTIFICATE OF SERVICE**

I certify that a copy of this document has been forwarded by electronic mail today to Plaintiff's counsel of record, Courtney Quish, Esq., BROMBERG & SUNSTEIN, LLP, 125 Summer Street, 11<sup>th</sup> Floor, Boston, Massachusetts 02110-1618.

Dated: October 18, 2006

  
Donald K. Piper